Revue québécoise de linguistique

Variations morphologiques, syntaxiques, sémantiques et Repérage d'Information sur le Web

Louisette Emirkanian et Emmanuel Chieze

Volume 32, numéro 1, 2003

TALN, Web et corpus

URI : https://id.erudit.org/iderudit/012247ar
DOI : https://doi.org/10.7202/012247ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Emirkanian, L. & Chieze, E. (2003). Variations morphologiques, syntaxiques, sémantiques et Repérage d'Information sur le Web. *Revue québécoise de linguistique*, *32*(1), 135–154. https://doi.org/10.7202/012247ar

Résumé de l'article

Le repérage d'information sur le Web présente des défis particuliers, en raison de la grande variété de domaines, genres et styles des documents (ce qui augmente les phénomènes de polysémie, d'homonymie et de synonymie), et des types de requêtes utilisées, en général très courtes. En conséquence, les résultats d'une recherche sont souvent très nombreux et peu pertinents. Il faut donc trouver des approches intermédiaires : nous avons étudié les résultats de cinq requêtes de base et de variantes obtenues par enrichissement morphologique et synonymique, dans le but d'identifier des pistes valables de reformulation de requêtes. Nous avons porté une attention particulière au lien syntaxique entre les termes de la requête dans les documents et à son rapport avec la pertinence de ces termes, et effectivement constaté que la prise en compte de ce lien devrait permettre d'augmenter la précision des requêtes sans trop nuire à leur rappel.

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal

VARIATIONS MORPHOLOGIQUES, SYNTAXIQUES, SÉMANTIQUES ET REPÉRAGE D'INFORMATION SUR LE WEB*

Louisette Emirkanian
Emmanuel Chieze
Université du Québec à Montréal

1. Introduction

1.1 Problématique générale du Repérage d'Information

Le Repérage d'Information (RI) sur le Web[1] est basé sur le modèle du sac de mots, les documents étant considérés comme des multiensembles de mots, et les requêtes comme des ensembles de mots. Plusieurs techniques sont associées à ce modèle pour rapprocher requêtes et documents : l'approche booléenne filtre les documents satisfaisant à une formule booléenne, ne tenant compte que de la présence ou non des termes de la requête dans les documents (ces derniers étant donc considérés comme des ensembles de mots), tandis que l'approche vectorielle classe les documents selon leur degré de ressemblance à la requête, en intégrant à cette formule de classement le nombre d'occurrences des termes de la requête dans les documents.

Ce modèle repose donc sur l'hypothèse simplificatrice implicite que les mots constituent des unités de sens atomiques et que le sens d'un document est la somme des sens des mots individuels. De plus, il ne vise à représenter que la dimension thématique de la pertinence (Cosijn et Ingwersen 2000), en excluant notamment les questions de domaine et de genre textuels, pourtant importantes sur le Web étant donné l'hétérogénéité de cette collection de documents.

Même en se restreignant à la dimension thématique de la pertinence, ce modèle possède de nombreuses limites. Premièrement, les mots dont il est question

* Nous tenons à remercier les évaluateurs pour leurs précieux commentaires et leurs suggestions.

1 Nous emploierons ici le terme *Web* pour désigner l'ensemble des documents interreliés auxquels nous accédons par Internet en utilisant le protocole http.

ici sont en réalité des graphies, ce qui implique que les variantes flexionnelles d'une même unité lexicographique sont considérées comme autant de mots indépendants, alors qu'une telle distinction n'est le plus souvent pas pertinente en RI (*voyage* et *voyages* devraient être traités de façon similaire en général). Deuxièmement, une graphie donnée n'est pas nécessairement associée à un sens unique, l'homonymie et la polysémie étant des phénomènes linguistiques très fréquents (le dessinateur Tibet n'a aucun rapport avec le Tibet). Troisièmement, le modèle de sac de mots ne rend pas compte de la dimension paradigmatique de la langue : des termes liés entre eux par des relations sémantiques ou conceptuelles (synonymie, hyperonymie, holonymie) seront considérés comme autant de termes indépendants (*marche* et *randonnée*). Quatrièmement, la dimension syntagmatique de la langue est absente de ce modèle, ce qui a au moins deux impacts : les concepts désignés par une expression ne sont pas explicitement pris en compte (une *lune de miel* n'a ainsi guère de rapport avec *la lune*), et l'articulation entre les concepts de la requête est également ignorée (une *marche au Tibet* est pourtant d'une autre nature qu'une *marche pour le Tibet*). En pratique cependant, les moteurs de recherche permettent de pallier partiellement cette limite en autorisant l'emploi d'expressions dans la requête, ces dernières étant utilisées comme filtre. Mais l'intégration complète des expressions au modèle vectoriel, à des fins de classement et non de simple filtrage, reste problématique (Perez-Carballo et Strzalkowski 2000). De plus, ce palliatif suppose que les utilisateurs du Web identifient eux-mêmes les expressions au sein des requêtes qu'ils écrivent, ce qui est très rarement le cas sur le Web (seulement 6 % des requêtes dans l'étude de Jansen, Spink et Saracevic 2000).

Il faut cependant noter que ces insuffisances sont parfois un avantage : ce modèle permet ainsi de repérer des documents sur un sujet donné sans qu'aucun lien explicite (de nature syntaxique ou sémantique) ne soit établi entre les termes de la requête, ce qui permet entre autres de traiter à l'identique certaines variations syntaxiques (*le voyage au Tibet*, *il voyage au Tibet*, *un voyage d'aventures au Tibet*), ou de repérer des documents dans lesquels il n'existe pas d'articulation explicite entre les termes de la requête.

1.2 Difficultés additionnelles du RI sur le Web

Le Web pose des défis particuliers, comparativement aux petites collections de documents plus classiquement utilisées en RI. Un premier défi est le caractère non contrôlé de la collection : n'importe qui peut ajouter, modifier et supprimer des documents sans que les index ne soient automatiquement mis à jour. En conséquence, des documents identifiés par un moteur de recherche

pourront ne plus exister au moment de la recherche. Des documents pourront exister en de multiples exemplaires (parfaitement identiques ou seulement similaires), ce qui entraîne une distorsion des résultats. Les documents sont enfin spécifiés dans divers formats informatiques, et sont surtout de qualité inégale, pouvant éventuellement contenir de nombreuses coquilles, ce qui a un impact sur le RI.

Mais le défi majeur du Web réside dans sa taille gigantesque, ce qui entraîne non seulement des difficultés informatiques (accroissement des ressources matérielles nécessaires au RI, optimisation des algorithmes utilisés), mais change la nature même du RI selon Blair 2002. Le nombre de sens associés à chaque mot ou expression croît avec la taille de la collection, et la multiplicité des domaines et genres couverts par le Web contribue largement au phénomène, ainsi que le caractère actuel de la collection, toujours en évolution, et incorporant ainsi continuellement des néologismes, de nouveaux noms propres et de nouveaux emplois de termes courants. Le nombre de documents dans les résultats des recherches est considérable, et l'utilisateur ne peut pas tous les parcourir. Sur le Web, les requêtes comprennent en moyenne deux termes (Jansen et Pooch 2001) : elles ne sont donc généralement pas suffisamment descriptives des besoins en information de l'utilisateur, et sont encore moins discriminantes. De plus, la désambiguïsation implicite des termes de la requête par leurs voisins, qui se produit pour des requêtes plus conséquentes, n'opère pas dans le cas de requêtes très courtes. Et ces dernières ne donnent en général aucun indice sur le domaine ou le genre des documents recherchés. En conséquence, il nous semble prioritaire de nous attaquer à la reformulation de requêtes dans le contexte du Web, en privilégiant la précision sur le rappel puisque les utilisateurs ne consultent généralement que la première page de résultats, soit au plus 10 documents.

2. Description générale du projet et présentation du corpus[2]

Jusqu'à récemment, les recherches dans le domaine de la RI et celles dans le domaine du traitement automatique des langues n'entretenaient que peu de liens, chaque domaine développant des outils spécifiques. Depuis peu, la convergence des deux domaines s'est faite et elle a porté ses fruits dans chacun d'eux (Jacquemin et coll. 2000, Jacquemin et Zweigenbaum 2000, Strzalkowski

2 Nous avons choisi d'utiliser le terme *corpus* par soucis de concision. Ce terme réfère ici à l'ensemble des documents retournés en réponse à un ensemble de requêtes exécutées pour étudier un besoin d'information donné.

1995, Strzalkowski et coll. 1999, Spärck Jones 1999, Woods et coll. 2000). Des travaux récents tendent à prouver que l'utilisation de techniques de TAL en morphologie, syntaxe et sémantique permet d'améliorer la performance des systèmes de RI tant au niveau de l'indexation qu'à celui du repérage : ces connaissances linguistiques sont utilisées pour le découpage en unités linguistiques, pour l'étiquetage, l'analyse en constituants (et l'indexation de syntagmes), la reconnaissance de termes complexes, la désambiguïsation en contexte et la reformulation de requêtes (Habert et Jacquemin 1993, Bourigault 1996, Dias et coll. 2000, Jacquemin et Tzoukermann 1999, Krovetz 1993, Namer 2000, Dal et Namer 2000, Bouillon et coll. 2000, Gaussier et coll. 2000).

Notre projet de recherche[3] a pour objectif d'évaluer l'apport des connaissances linguistiques à l'amélioration du rappel et de la précision, en mettant à jour des pistes pour la spécification de mécanismes de reformulation de requêtes facilitant la recherche d'information.

À partir de 5 besoins en information de type MOT1 MOT2, «FUITE DES CERVEAUX» ÉTATS-UNIS (où *fuite des cerveaux* constitue le MOT1), VOYAGE TIBET, VOL LUNE, MISSION ESPACE, PROMENADE PARIS (soit les corpus FUITE DES CERVEAUX, TIBET, LUNE, ESPACE et TIBET), nous avons constitué 5 corpus. On peut se référer à l'article de Fouqueré et Issac ici même pour plus d'informations sur l'extraction des données du Web et la constitution du corpus. Plusieurs requêtes de la forme MOT1 (NEAR préposition) NEAR MOT2 ont été exécutées pour constituer chacun des corpus, avec des variations morphologiques et sémantiques sur MOT1 et MOT2 ainsi que des variations sur la préposition[4]; par exemple, dans le cas de la requête VOYAGE TIBET, *voyage* est remplacé par *voyages*, *séjour*, *voyager*, *trek*, etc. et *Tibet* par *tibétain*, *tibétaines*, etc. Dans le cas où le MOT2 est un adjectif, la préposition pourra être présente ou non (*promenade parisienne* ou *promenade dans les arrondissements parisiens*). Cette combinaison (MOT1, préposition, MOT2) a généré 1211 requêtes différentes que nous avons lancées sur le Web. Elles nous ont permis de récupérer environ 34000 pages Web.

La pertinence informationnelle est la pertinence classique du RI, qui identifie les documents répondant complètement ou partiellement au besoin de l'utilisateur, sans nécessairement établir de liens avec la formulation particulière de la requête. La pertinence thématique est quant à elle rattachée à un passage contenant les termes de la requête, et non au document dans son ensemble, et juge si ces termes sont employés dans un sens compatible avec le besoin en

3 Ce projet a été financé par la coopération franco-québécoise (ministère des Relations internationales du Québec et ministère des Affaires étrangères de la France).

4 Nous avons utilisé une liste fixe de 31 prépositions indépendante des besoins d'information étudiés.

information et s'ils constituent l'objet du discours rattaché au passage. Pour chacun de nos corpus, les besoins en information ont été définis comme suit. Pour le corpus FUITE DES CERVEAUX, répondaient à notre besoin en information les documents traitant de la fuite des cerveaux vers les États-Unis; pour le corpus TIBET, tous les documents qui donnaient des informations pratiques pour l'organisation d'un voyage touristique au Tibet; pour le corpus ESPACE, tous les documents donnant des informations scientifiques sur les missions dans l'espace intersidéral; pour le corpus LUNE, tous les documents donnant des informations scientifiques ou des détails sur les vols (effectués ou à venir) vers la lune; enfin pour le corpus PARIS, tous les documents donnant des informations pratiques permettant d'organiser des promenades dans Paris.

Nous nous attacherons ici à l'étude de la pertinence thématique par rapport aux contextes d'emplois des termes de la requête, aux variations morphologiques et sémantiques ainsi qu'au lien syntaxique entre les termes.

Nous avons constitué un sous-ensemble représentatif pour chacun des 5 besoins en information. Nous n'avons retenu que les occurrences dans lesquelles MOT1 préposition MOT2 apparaissent dans l'ordre de la requête et au sein de la même phrase. Nous avons ainsi obtenu 3225 occurrences réparties dans 2030 documents distincts. 62 % d'entre elles ont été jugées thématiquement satisfaisantes. Elles se répartissent comme suit dans les cinq corpus (Tableau 1).

Tableau 1
Évaluation de la pertinence thématique sur l'ensemble du corpus

CORPUS	NOMBRE D'OCCURRENCES	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
FUITE DES CERVEAUX	119	88 %
TIBET	1061	62 %
LUNE	782	70 %
ESPACE	551	68 %
PARIS	712	44 %
TOTAL	**3225**	**62 %**

Après avoir examiné les contextes d'emplois des termes de la requête, nous parlerons de l'impact sur la pertinence thématique des variations morphologiques et sémantiques, et de la présence d'un lien syntaxique entre les termes.

Ces premières analyses vont nous aider à mettre à jour des pistes nous permettant de filtrer les résultats obtenus (augmentation de la précision) ou de les élargir (augmentation du rappel), et ainsi de contrôler la reformulation des requêtes.

3. Le contexte d'emploi des termes de la requête

La prise en compte du contexte d'emploi vise à augmenter la précision. Nous avons systématiquement examiné les différents éléments gravitant autour de MOT1 et de MOT2, ceux qui les précèdent ou les suivent. Nous avons isolé trois éléments jouant un rôle important, tout au moins dans notre corpus, et permettant d'éliminer certaines occurrences thématiquement incorrectes : l'orthographe, les déterminants et l'appartenance de l'un des termes à une unité lexicale complexe.

3.1 L'orthographe

La prise en compte de certains signes auxiliaires, le trait d'union par exemple, peut être discriminante. Dans le cas du corpus PARIS, les séquences dans lesquelles *Paris* apparaît précédé ou suivi d'un trait d'union (*Paris-Dakar*, *Bordeaux-Paris*, etc.) ne sont pas liées au thème de la requête. Il en est de même dans les corpus LUNE et ESPACE lorsque le MOT1 *volant* est précédé du trait d'union et qu'il entre dans un composé du type *cerf-volant*. Cependant, l'expérience montre que les règles régissant l'emploi des signes auxiliaires ne sont pas toujours respectées dans les documents très variés provenant du Web, pas plus d'ailleurs que celles concernant l'emploi des majuscules qui peuvent également jouer un rôle, par exemple, dans la distinction entre le dérivé adjectival et le dérivé nominal de *Tibet* (*tibétain*/*Tibétain*). Les signes de ponctuation permettent également de rejeter des séquences non pertinentes au niveau thématique; dans le cas où *vol* est suivi d'un point et le plus souvent d'un chiffre (*vol.*), on est en présence de l'abréviation de *volume*. Enfin, les accents peuvent jouer un rôle dans la discrimination de certains termes : *la marche/le marché*. Dans le cas de cette paire, la prise en compte du genre du déterminant est également intéressante.

3.2 Les déterminants

Les déterminants, dans plusieurs cas, permettent d'éliminer certaines occurrences incorrectes au niveau thématique. La présence du déterminant (contracté avec la préposition ou non) différencie certaines formes nominales homographes des formes verbales recherchées (participe présent ou participe passé). On peut par exemple rejeter les séquences du type *le volant*[5]/*au volant*,

5 Comme nous l'a fait remarquer un évaluateur, précisons que *le*, pronom, dans le contexte *(en) le volant*, serait également éliminé.

le marché/*du marché* et conserver celles où *volant* et *marché* sont respectivement des formes des verbes *voler* et *marcher*. Inversement, pour le MOT2, dans les corpus TIBET, ESPACE et LUNE (un nom propre et deux quasi-noms propres), on s'attend à trouver un déterminant, pouvant être contracté dans le corpus TIBET. Ainsi, les cas où le MOT2 apparaît sans déterminant peuvent être exclus : *par Tibet*, *avec Tibet* (nom de famille); dans le corpus LUNE chaque fois que *de lune* apparaît (*clair de lune)*, nous sommes en présence d'occurrences non pertinentes. Quant au mot *espace*, sur les 50 cas où il apparaît sans déterminant, deux seulement sont thématiquement corrects. Le type du déterminant peut également permettre le filtrage des résultats. Dans le cas où le MOT2 est un terme ayant un référent unique, un quasi-nom propre (*la lune*, *l'espace*), seul le déterminant défini peut apparaître. Ainsi, les cas où *lune* et *espace* sont précédés de déterminants possessifs, démonstratifs ou indéfinis (*votre lune de miel*, *cet espace*, *un espace de séjour*) peuvent-ils être éliminés.

3.3 L'appartenance de l'un des termes à une unité lexicale complexe

Dans de nombreux cas, le MOT1 ou le MOT2 se trouvent en cooccurrence avec des termes spécifiques avec lesquels ils forment une unité lexicale complexe.

Dans la majorité des cas où le MOT1 (dans le corpus PARIS) est *déplacement* (en variation synonymique avec *promenade*), les résultats ne sont pas satisfaisants chaque fois que ce mot se trouve dans des séquences du type *frais de déplacement*, *plan de déplacement*, *politique des déplacements*, *problème des déplacements*. Dans ce même corpus, lorsque *marché* est suivi d'un syntagme prépositionnel du type *du logement*, *de l'immobilier*, *de la bourse*, *du travail*, le document ne porte pas sur le fait de marcher dans Paris. C'est surtout pour les corpus ESPACE et LUNE que la prise en compte du contexte d'emploi peut être utile dans la mesure où les mots *lune* et *espace* entrent souvent dans la composition d'unités lexicales complexes, les plus fréquentes étant *lune de miel*, *clair de lune*, *nouvelle lune*, *pierre de lune*, *pleine lune*, *croissant de lune*, et *espace aérien* ou encore *espace Schengen*.

Le contexte d'emploi des termes d'une requête peut donc constituer un critère pour filtrer les résultats et reformuler les requêtes après avoir proposé à l'utilisateur les contextes les plus fréquents dans lesquels les termes de sa requête apparaissent. Il est cependant difficile d'appliquer dans tous les cas ce genre de filtre. Par exemple, l'unité complexe *agence de voyage* (dans le corpus TIBET) ne donne pas que des résultats non pertinents. Néanmoins, on remarque que souvent plusieurs critères se complètent. Par exemple, on peut éliminer *cerf-volant* d'une part par la présence du trait d'union, d'autre part par le fait

que *volant* fait partie d'une unité lexicale complexe. Les séquences du type *clair de lune* peuvent être rejetées non seulement par le fait que les termes apparaissent souvent en cooccurrence, mais aussi par l'absence de déterminant pour *lune*. Quant aux séquences du type *marché de la bourse*, elles peuvent être éliminées parce qu'elles constituent une unité lexicale complexe, mais aussi par le fait que *marché* est en général précédé d'un déterminant.

4. Variations synonymiques et morphologiques des termes de la requête

Alors que l'étude du contexte d'emploi des termes de la requête permet d'accroître la précision, les variations synonymiques et morphologiques ont pour objectif d'améliorer le rappel.

4.1 Variations synonymiques

Dans nos requêtes, nous avons intégré des variations surtout sur le MOT1. Dans le corpus FUITE DES CERVEAUX, cette expression est en variation *exode/exil des chercheurs/cerveaux*, *fuite des chercheurs/cerveaux*. Pour le corpus TIBET, parallèlement à *voyage*, nous avons *trek*, *aventure*, *marche*, etc.; pour ESPACE, *mission* est en variation avec *vol* et *séjour*. *Vol* est remplacé par *voyage* et *mission* dans le corpus LUNE. Enfin, dans le corpus PARIS, *promenade*, *balade*, *aventure*, *vadrouille*, *pérégrination*, *déplacement*, etc. constituent des MOT1 possibles.

Les résultats nous montrent que la sélection de termes trop distants diminue considérablement la précision sans réellement augmenter le rappel. Certains termes, tels que *pérégrination*, *vagabondage*, etc., sont trop peu employés pour justifier leur apport à la requête d'autant plus qu'ils appartiennent à un registre susceptible d'identifier d'autres types de documents que ceux qui sont recherchés.

D'autres mots tels que *aventure* ou *errer* (et ses dérivés) font dévier la requête; avec *aventure* (dans le corpus PARIS), on récupère un grand nombre de documents faisant référence au titre d'un livre d'aventures édité à Paris. Dans ce corpus, si *promenade(s)* et *balade(s)* donnent de bons résultats, les autres MOT1, même s'ils augmentent le rappel, diminuent considérablement la précision (70 % de cas non désirés). *Déplacer* (et ses dérivés), dans les corpus PARIS et TIBET, constitue un mauvais choix puisque ce verbe semble plus souvent dénoter une transition qu'un procès (Paris est le lieu de destination du déplacement). Dans le corpus LUNE, alors que l'emploi de *mission(s)* et de *vol(s)* donne de bons résultats (81 % de séquences thématiquement correctes), celui de *voyage(s)* diminue considérablement la précision (seulement 33 % de séquences satisfaisantes),

ce mot se trouvant le plus souvent dans des séquences du type *Voyage dans la Lune,* qui renvoient à une œuvre artistique. Pour le corpus TIBET, seuls *trek* et son dérivé *trekking* permettent d'améliorer à la fois le rappel et la précision (84 % de séquences thématiquement correctes). Tous les autres MOT1 en variation avec *voyage* ont permis de rapatrier un grand nombre d'occurrences, mais 67 % d'entre elles ne correspondent pas au sens désiré. En particulier, le mot *marche* constitue un mauvais choix. Outre le fait qu'il permet de rapatrier *(le) marché*, on le retrouve très souvent dans des séquences du type *marche pour le Tibet,* qui font dévier le sens de la requête de base. Sur les 95 cas où *marche(s)* constitue le MOT1, seuls 4 cas sont thématiquement corrects.

Il semble donc difficile d'élargir a priori la requête par l'ajout de termes en relation sémantique ou conceptuelle avec les éléments de base : on risque en effet de faire dévier le sens de la requête de base. On pourrait cependant proposer à l'utilisateur, s'il y a lieu, des séquences où certains éléments sont en relation paradigmatique; par exemple, dans le corpus TIBET, un nombre significatif de documents contiennent à la fois *voyage au Tibet* et *trek/trekking au Tibet*.

4.2 Variations morphologiques

Les variations morphologiques, tout comme les variations synonymiques, ont pour but d'améliorer le rappel. Ces variations portent sur le MOT1 et le MOT2, à l'exception du corpus FUITE DES CERVEAUX, où aucune variation morphologique n'a été effectuée[6]. Avant d'aborder l'apport de la prise en compte de la morphologie dérivationnelle, précisons que la morphologie flexionnelle semble avoir peu d'incidence. Le pourcentage d'occurrences thématiquement correctes est le même, sauf dans un cas : celui de *marches*, où le pluriel fait référence à un sens différent (*marches sino-tibétaines*, *marches du Tibet*) qui fait dévier le sens de la requête de base dans 18 cas sur 19.

La morphologie dérivationnelle s'avère plus intéressante. Pour le MOT1, nous avons des dérivés nominaux et verbaux, et pour le MOT2, des dérivés adjectivaux.

4.2.1 Morphologie dérivationnelle nominale et verbale (MOT1)

Le premier problème concerne la formalisation de la dérivation dans la requête, l'objectif étant de rapatrier les dérivés du mot de base, et seulement ceux-là. L'utilisation de l'astérisque est dans certains cas peu souhaitable.

6 L'emploi métaphorique de *fuite* dans *fuite des cerveaux* nécessite le pluriel. Après une vérification a posteriori sur le Web, nous n'avons trouvé qu'une occurrence de *fuite d'un cerveau*.

Par exemple, *vol**[7] entraîne le rapatriement de documents contenant des non dérivés (*volet*, *volonté*, etc.). Pour les quatre corpus concernés par la variation morphologique, sur les 1032 occurrences supplémentaires obtenues, seules 610 contiennent un dérivé nominal ou verbal du mot de base.

Parmi les dérivés nominaux des mots de base (159 pour l'ensemble du corpus), 45 % seulement se trouvent dans des séquences thématiquement satisfaisantes. En effet, certains font dévier le sens de la requête : *missionnaire* et *aventurier* en sont des exemples; *marcheurs* également, qu'on retrouve le plus souvent dans le sens de «sportifs» ou de «participants à une manifestation». Le seul cas qui nous permette d'augmenter le rappel et la précision est celui de *trekking*, qui apparaît chaque fois dans des séquences liées au thème de la requête.

Les résultats sont sensiblement meilleurs avec les dérivés verbaux puisque 54 % des occurrences sont satisfaisantes. En revanche, le rappel est faible puisque les verbes ne représentent que 15 % de l'ensemble du corpus (Tableau 2).

Tableau 2
Évaluation de la pertinence thématique avec les dérivés verbaux.

CORPUS	NOMBRE D'OCCURRENCES	NOMBRE DE VERBES	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
TIBET	1 061	169	49 %
LUNE	782	46	48 %
ESPACE	551	41	59 %
PARIS	712	195	60 %
TOTAL	**3 106**	**451**	**54 %**

Il est intéressant de remarquer que, pour le corpus PARIS, les verbes donnent de meilleurs résultats que l'ensemble des noms et de leurs dérivés. Comme nous avons pu le constater dans le Tableau 1, nous avons obtenu pour ce corpus 44 % de séquences thématiquement correctes; or, lorsque MOT1 est un verbe, nous en obtenons 60 %. La plupart de ces séquences apparaissent dans des documents appartenant plutôt au genre littéraire dans lesquels ces verbes sont conjugués.

Il est à noter que les verbes sont rarement utilisés en repérage de l'information. Lorsqu'ils le sont, cela est souvent sans grand succès, comme c'est le cas ici, à l'exception, nous venons de le voir du corpus PARIS. En effet, le

7 L'astérisque, avec Altavista, remplace de zéro à cinq caractères positionnés en fin de mot. Par exemple, *vol** identifiera les termes *vol*, *vols*, *volume* mais pas *volumineux*.

repérage de l'information s'attache à identifier des documents sur un thème donné, défini par un concept (un ou plusieurs termes : *promenade à Paris*), ou par des liens entre des concepts ((*impacts de la*) *pollution (sur la) politique des transports à Paris*). Des substantifs sont indispensables pour décrire ces concepts en tant qu'objets du discours. Le verbe quant à lui établit une relation en général transitoire, ponctuelle, voire anecdotique entre plusieurs concepts; le sens qu'il véhicule ne constitue pas l'objet du discours (*Venue par hasard, je me baladais à Paris et passant devant un théâtre...*), et le fait de transformer, dans nos requêtes, un nom en verbe supprime du coup un concept. De plus, les verbes (*balader*) sont généralement moins couramment employés que les substantifs correspondants (*balade*). La difficulté de les identifier à cause de la flexion en particulier, et leur peu d'apport comme rappel ne semblent pas justifier leur emploi en repérage de l'information. Cependant, l'infinitif constitue peut-être une exception, de par ses emplois quasi substantivaux, notamment dans les titres (*moyens de se déplacer dans Paris; se déplacer à Paris*).

4.2.2 Morphologie dérivationnelle adjectivale (MOT2)

Les dérivés adjectivaux pour le MOT2 sont légèrement plus nombreux que le sont les dérivés nominaux et verbaux pour le MOT1 : ils sont en effet présents dans 24 % des occurrences pour les quatre corpus concernés. De plus, 77 % d'entre elles se trouvent dans des séquences reliées au thème. Ils semblent donc intéressants au niveau du rappel et de la précision. Il faut cependant observer les différences selon les corpus comme le montre le Tableau 3.

Tableau 3
Évaluation de la pertinence avec les dérivés adjectivaux

CORPUS	NOMBRE D'OCCURRENCES	NOMBRE D'ADJECTIFS	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
TIBET	1 061	197	32 %
LUNE	782	277	95 %
ESPACE	551	224	99 %
PARIS	712	37	46 %
TOTAL	**3 106**	**735**	**77 %**

Les corpus TIBET et PARIS se comportent différemment des corpus ESPACE et LUNE, aussi bien au niveau du rappel qu'à celui de la précision. Les adjectifs *lunaire* et *spatial*, dérivés de deux quasi-noms propres, sont très nombreux. Le premier représente 35 % des MOT2 du corpus LUNE, le second, 41 % des

MOT2 du corpus ESPACE. Pour ce qui est de *tibétain* et de *parisien*, en revanche, le rappel est moindre. On trouvera plus fréquemment des *promenades dans Paris* ou *dans les rues de Paris* que *des promenades parisiennes* ou *dans les rues parisiennes*.

Alors que dans le cas de *parisien* et de *tibétain,* on peut être confronté au risque d'élargir la signification du terme de base, au contraire, dans le cas de *lunaire* et de *spatial*, il y a plutôt une restriction des sens admissibles. En effet, *lunaire* ne se rapporte qu'à la lune au sens propre et élimine les autres *lunes*; quant à *spatial*, il ne peut faire référence qu'à l'espace interplanétaire. Ces deux adjectifs peuvent se trouver dans la séquence 'MOT1 adjectif', comme dans *mission lunaire*, *mission spatiale*, *vol spatial*; ou encore, l'adjectif peut modifier un autre élément que le MOT1 : *premier vol d'une sonde lunaire*, par exemple. Pour ces deux corpus, l'adjectif a donc une double incidence, sur le rappel et la précision. Il est intéressant de noter que si 95 % des cas où l'adjectif *lunaire* apparaît sont thématiquement corrects, le MOT2 *lune* donne de moins bons résultats (56 %); on retrouve la même différence entre *spatial* (99 % de cas thématiquement corrects) et *espace* (46 % de cas thématiquement corrects).

Alors que les dérivés verbaux et nominaux améliorent le rappel au détriment de la précision, les dérivés adjectivaux semblent plus intéressants surtout dans le cas où ils permettent de restreindre le sens d'un des termes de la requête de base.

5. Étude du lien syntaxique entre les termes de la requête

Jusqu'ici, nos observations ne concernaient que les termes individuels de la requête. Nous allons, dans ce qui suit, examiner l'importance du lien syntaxique entre les termes pour l'amélioration de la précision.

5.1 Catégorisation des liens syntaxiques observés

Nous avons réparti les occurrences obtenues en six catégories de liens syntaxiques; elles sont liées au fait, d'une part, que nous autorisons différentes catégories pour MOT1 ou MOT2, et, d'autre part, que nos requêtes contiennent l'élément NEAR.

Voici quelques exemples d'occurrences obtenues pour chacune des constructions; nous avons fait précéder d'un astérisque les cas thématiquement non pertinents.

5.1.1 N1-SP

Le MOT2 est dans un syntagme prépositionnel rattaché au MOT1 :

(1) … nous revoyons également les grandes lignes de l'entraînement rigoureux qui a fait de Julie Payette une astronaute prête à accomplir sa première mission dans l'espace.
http://radio-canada.ca/tv/decouverte/semaine/990516cs.html

(2) *Lorsque les conditions météorologiques sont inférieures aux conditions minimales requises pour les vols VFR en espace contrôlé, …
http://www.cavok-fr.com/reglementation/vfr-spec.htm

(3) Pour ce 7e voyage en quinze ans en terre tibétaine et afin d'éviter d'éventuels tracas de visa, nous sommes montés par la Chine occidentale … **http://www.amis-tibet.lu/t_info17/TibetToday17.htm**

5.1.2 N1-ADJ

Le MOT2 est un adjectif qui modifie directement le MOT1 :

(4) En effet, il avait développé pendant les quatorze ans de son séjour tibétain un grand attachement pour le Tibet …
http://www.tibet-info.net/info/tibet_info/2000/31_12.html

(5) *Partie le 9 juillet de Nice, la «marche transalpine tibétaine», composée d'une cinquantaine de marcheurs qui protestent contre ...
http://www.tibet-info.net/info/tibet_info/2000/31_08.html

5.1.3 P-N2

Le MOT2 est dans un syntagme prépositionnel non relié au MOT1 :

(6) Au cours du vol de Gemini 4, Edward White effectua lui aussi une sortie dans l'espace, le 3 juin 1965.
http://www.multimania.com/msegret/laconqu.htm

(7) *L'Agence universitaire a pour mission de contribuer à la construction d'un espace universitaire francophone …
http://www.clf.gouv.qc.ca/autres.htm

5.1.4 AUCUN LIEN

Les trois termes MOT1, préposition, MOT2 ne sont pas reliés syntaxiquement :

(8) Un voyage très complet qui permet de découvrir à la fois le Tibet central et l'ancien royaume de Gugé.
http://tirawa.com/voyages/himalaya/tibet/tibet-703-lhassa-mont-kailash/index.html

(9) *Elle aura comme mission principale le déploiement du satellite d'observatoire X (rayonnements X) appelé Chandra (Chandra veut dire Lune ou lumière en sanskrit).
http://www-lm2s.univ-troyes.fr/~chatelet/divers/astro_arch.html

5.1.5 V-SP

Le MOT1 est un verbe; le MOT2 est dans un syntagme prépositionnel rattaché au verbe :

(10) Voyager au Tibet est une expérience incroyable, mais cela nécessite une bonne condition physique.
http://www.multimania.com/portesaventure/tibetfr.htm

(11) *… on pourrait apercevoir de vieilles bonne-femmes voler sur leur balai devant la lune. **http://www.codeco.qc.ca/magnet/9910/societe3.htm**

5.1.6 V-SN

Le MOT1 est un verbe; le MOT2 est dans un syntagme nominal argument du verbe :

(12) Ils ont parcouru sacs au dos, l'Inde, le Tibet, le Népal et la Chine.
http://www.lafirme.com/site/bullreg/chroniqu/lecture/recit.html

(13) *Pour s'y rendre nous traversons des camps de réfugiés tibétains;
http://www.geocities.com/TheTropics/Coast/2384/voyag1c1.htm

Comme ces exemples le montrent, dans les différentes structures, le MOT1 peut être modifié, ou encore le MOT2 peut être un modifieur. Nous reviendrons sur ces cas au point 5.2.2.2.

5.2 Analyse des résultats

Les résultats pour chacune des structures sont consignés dans le Tableau 4.

Tableau 4
Évaluation de la pertinence thématique selon les structures

STRUCTURES	NOMBRE D'OCCURRENCES	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
N1-SP	1152	77 %
N1-ADJ	282	98 %
P-N2	898	46 %
AUCUN LIEN	632	33 %
V-SP	220	80 %
V-SN	41	95 %
TOTAL	**3225**	**62 %**

On constate, comme on pouvait s'y attendre, que les liens N1-ADJ, V-SN, V-SP et N1-SP donnent de meilleurs résultats que les cas P-N2 ou encore ceux où il n'y a aucun lien entre les trois termes. Nous allons considérer d'abord les cas N1-ADJ, V-SP, V-SN et AUCUN LIEN, ensuite les cas N1-SP et P-N2.

5.2.1 Analyse des cas N1-ADJ, V-SP, V-SN et AUCUN LIEN

Dans le cas du lien N1-ADJ, comme on l'a déjà dit, les résultats varient selon le corpus concerné : pour les corpus LUNE et ESPACE, l'apport de l'adjectif est intéressant aussi bien pour le rappel que pour la précision. Pour ce qui est des liens V-SP et V-SN, la précision est nettement supérieure à la moyenne (82 % au lieu de 62 %), mais, rappelons-le, les cas où le MOT1 est un verbe sont rares (15 %). Évidemment, lorsque les termes MOT1, préposition et MOT2 n'ont aucun lien syntaxique, les résultats sont peu satisfaisants au niveau thématique. Cependant, il y a lieu dans ce cas également de considérer chacun des corpus isolément. Par exemple, dans le cas des corpus LUNE et ESPACE, les résultats sont meilleurs que dans les autres corpus au niveau de la précision (respectivement 65 % et 56 % de séquences thématiquement correctes dans la catégorie AUCUN LIEN) surtout lorsque les MOT2 sont les adjectifs. *Lunaire* et *spatial*, en effet, sont presque dans tous les cas en cooccurrence avec les substantifs *module*, *sol*, *sonde*, *station*, *navette*. On note une fois de plus l'importance de l'emploi de l'adjectif dans ces deux corpus.

5.2.2 Analyse des cas N1-SP et P-N2

Examinons maintenant les liens N1-SP et P-N2. Le cas N1-SP représente 36 % des occurrences du corpus. Pour la précision, 77 % d'entre elles sont thématiquement correctes; le lien P-N2 représente quant à lui 28 % des occurrences avec 46 % de séquences satisfaisantes. On aurait pu s'attendre, dans le cas du lien N1-SP, à un pourcentage plus élevé de séquences thématiquement correctes; les variations synonymiques qui font dévier la requête de base ainsi que les prépositions permettent de rendre compte de la majorité des cas non satisfaisants (*Voyage dans la lune* et *marche pour le Tibet* sont des exemples représentatifs de ces cas). Aussi n'examinerons-nous dans les analyses suivantes que les cas où MOT1 et MOT2 ne font en aucun cas dévier le sens de la requête de base : *voyage(s)/trek/trekking* (préposition) *Tibet*, *promenade(s)/balade(s)* (préposition) *Paris*, *vol(s)/mission(s)* (préposition) *lune*, *vol(s)/mission(s)* (préposition) *espace*, *fuite/exode des cerveaux/chercheurs* (préposition) *États-Unis*. Nous considèrerons dans un premier temps les cas où MOT1 préposition MOT2 forment une séquence continue, le syntagme prépositionnel déterminant le MOT1 (*voyage au Tibet*) et dans un second temps, les cas où les trois termes forment une séquence discontinue.

5.2.2.1 Les séquences continues

Dans chacun de nos corpus, le pourcentage de séquences thématiquement correctes est très élevé comme le montre le Tableau 5.

Tableau 5
Évaluation de la pertinence thématique pour la séquence continue
MOT1 préposition MOT2 (quelle que soit la préposition)

SÉQUENCES CONTINUES	NOMBRE D'OCCURRENCES	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
Voyage(s) /trek/trekking (préposition) *Tibet*	238	99 %
Mission(s)/vol(s) (préposition) *espace*	47	87 %
Vol(s)/mission(s) (préposition) *lune*	43	100 %
Promenade(s)/balade(s) (préposition) *Paris*	59	97 %
Fuite/exode des cerveaux/chercheurs (préposition) *États-Unis*	33	100 %
TOTAL	**420**	**97 %**

Il est à noter que pour les différents corpus, dans 90 % des cas, les prépositions employées dans les séquences MOT1 préposition MOT2 sont bien sûr les suivantes : pour le corpus TIBET *au*, pour le corpus ESPACE *dans*, pour le corpus LUNE *vers* et *sur*, pour le corpus PARIS *dans*, et pour le corpus FUITE DES CERVEAUX *vers* et *aux*. Nous allons maintenant examiner les cas discontinus.

5.2.2.2 Les séquences discontinues

Nous avons distingué, rappelons-le, deux types de séquences discontinues, celles où le MOT1 est modifié et celles où le MOT2 est un modifieur. Il faut noter que les cas où le MOT2 est un modifieur sont rares (puisque nous n'avons pas considéré ici le cas où le MOT2 est un adjectif); parallèlement à ce faible rappel, nous perdrions beaucoup au niveau de la précision si nous tenions compte de ces cas dont le schéma MOT1 … préposition … MOT2 s'apparente à celui où les termes n'ont aucun lien entre eux et où le pourcentage de séquences thématiquement correctes est très faible. En effet, la prise en compte des séquences discontinues MOT1 … préposition … MOT2 nous permet de récupérer 152 séquences supplémentaires dont seulement 47 % sont thématiquement correctes.

Nous nous attacherons donc à considérer uniquement les cas où le MOT1 est modifié comme dans *mission habitée vers la Lune*. Il s'agit donc de la séquence MOT1 … préposition MOT2. Ce cas s'apparente à celui de la structure P-N2 où le SP ne modifie pas le MOT1 comme dans l'exemple suivant :

(14) Si vous souhaitez vous associer à tout ou partie d'un voyage touristique entre le 20 mai et le 15 juin, au Népal dont 15 jours au Tibet, nous contacter rapidement par e-mail. **http://www.tibet.fr/mars2000.htm**

La question qui se pose alors est de savoir s'il est pertinent en termes de rappel et de précision de tenir compte de ces cas, c'est-à-dire d'autoriser à côté des séquences continues des séquences discontinues du type MOT1 … préposition MOT2. En plus des 420 occurrences continues du type *voyage* (préposition) *Tibet*, nous obtenons 472 occurrences du type *voyage* ... (préposition) *Tibet* avec un lien syntaxique de détermination ou non entre le N1 et le SP. 75 % d'entre elles sont thématiquement pertinentes. Dès lors qu'on affine la recherche et qu'on note la préposition la plus couramment employée pour chacun des corpus (*voyage(s) ... au Tibet*, *mission(s)/vol(s) ... vers/sur la lune*, *promenade(s)/balade(s) ... dans Paris*, etc.), on récupère, en plus des 420 occurrences continues, 274 occurrences, dont 85 %[8] sont satisfaisantes (Tableau 6).

8 Le pourcentage serait plus élevé si nous avions éliminé au préalable les cas de *espace aérien* qui représentent la quasi-totalité des cas non satisfaisants dans le corpus ESPACE.

Tableau 6
Évaluation de la pertinence thématique
pour la séquence discontinue MOT1…préposition MOT2
(avec la préposition adéquate)

SÉQUENCES DISCONTINUES	NOMBRE D'OCCURRENCES	PROPORTION D'OCCURRENCES THÉMATIQUEMENT CORRECTES
Voyage(s) /trek/trekking ... au *Tibet*	88	89 %
Mission(s)/vol(s) ... dans l'*espace*	57	61 %
Vol(s)/mission(s) ... vers/sur la *lune*	84	90 %
Promenade(s)/balade(s)... dans *Paris*	9	100 %
Fuite/exode des cerveaux/chercheurs ... aux/vers les *États-Unis*	36	97 %
TOTAL	**274**	**85 %**

Les résultats obtenus montrent que la prise en compte de ces séquences discontinues, au moment de la reformulation de la requête, permet d'améliorer le rappel sans nuire à la précision.

6. conclusion

Même si les variations synonymiques et morphologiques peuvent dans certains cas améliorer le rappel, cela se produit en général au détriment de la précision soit parce que les termes résultants font dévier le sens de la requête, soit parce qu'ils entraînent le rapatriement de mots qui n'ont aucun rapport avec les termes de base. La dérivation adjectivale semble constituer un cas à part, en particulier dans le cas où elle permet de restreindre les sens admissibles. La prise en compte du contexte d'emploi et celle de la structure présente un intérêt certain au niveau de la précision, mais il semble difficile de prévoir a priori les filtres morphologiques et syntaxiques appropriés à un besoin d'information particulier. On pourrait détecter automatiquement et proposer à l'utilisateur les contextes les plus fréquents dans lesquels se trouvent les termes de la requête qu'il a soumise et, à partir de ses choix, filtrer les résultats. Ces contextes prendraient également en compte les prépositions, et les requêtes pourraient être reformulées sous la forme d'une séquence continue *mission dans l'espace*, *vol vers la lune* ou sous la forme d'une séquence discontinue *mission ... dans l'espace*, *vol ... vers la lune* (*mission* NEAR *«dans l'espace»*, *vol* NEAR *«vers la lune»*). En effet, la prise en compte des cas discontinus est

intéressante puisqu'elle améliore considérablement le rappel sans porter atteinte à la précision; dans trois de nos corpus, en particulier, on obtient plus d'occurrences discontinues que d'occurrences dans lesquelles le SP modifie directement le MOT1.

Références

BLAIR, D. C. 2002 «The challenge of commercial document retrieval, Part I: Major issues,and a framework based on search exhaustivity, determinacy of representation and document collection size», *Information Processing and Management* 38 : 273-291

BOUILLON, P., C. FABRE, P. SÉBILLOT et L. JACQMIN 2000 «Apprentissage de ressources lexicales pour l'extension de requêtes», *TAL* 41-2 : 367-393.

BOURIGAULT, D. 1996 «Lexter, a Natural Language Processing Tool for Terminology Extraction», *Proceedings of the 7th EURALEX International Congress*, pp. 771-779.

COSIJN, E. et P. INGWERSEN 2000 «Dimensions of relevance», *Information Processing and Management* 36 : 533-550.

DAL, G. et F. NAMER 2000 «Génération et analyse automatiques de ressources lexicales construites utilisables en recherche d'informations», *TAL* 41-2 : 423-446.

DIAS, G., S. GUILLORÉ, J.-Cl. BASSANO et J. GABRIEL PEREIRA LOPES 2000 «Extraction automatique d'unités lexicales complexes : un enjeu fondamental pour la recherche documentaire», *TAL* 41-2 : 447-472.

FOUQUERÉ, Ch. et F. ISSAC 2002 «Pertinence thématique de variations de requêtes», Communication au colloque TALN, Corpus et Web (Saint-Denis, France), texte ici même sous le titre «Corpus issus du Web : constitution et analyse informationnelle», *Revue québécoise de linguistique* 32-1.

GAUSSIER, E., G. GREFENSTETTE, D. HULL et Cl. ROUX 2000 «Recherche d'information en français et traitement automatique des langues», *TAL* 41-2 : 473-493.

HABERT, B. et Ch. JACQUEMIN 1993 «Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques», *TAL* 34-2 : 5-42.

JACQUEMIN, Ch. et coll. 2000 *Traitement automatique des langues pour la recherche d'information*, *TAL*, vol. 41, nº 2.

JACQUEMIN, Ch. et E. TZOUKERMANN 1999 «NLP for term variant extraction : Synergy of morphology, lexicon and syntax», dans Strzalkowski, T. et coll. 1999 pp. 25-74.

JACQUEMIN, Ch. et P. ZWEIGENBAUM 2000 «Traitement automatique des langues pour l'accès au contenu des documents», dans Le Maître J., J. Charlet, C. Garbay et coll. *Le document en sciences du traitement de l'information,* Toulouse, Cepadues, pp.71-109.

JANSEN, B. J. et U. POOCH 2001 «A Review of Web Searching Studies and a Framework for Future Research», *Journal of the American Society for Information Science and Technology* 52-3: 235-246.

JANSEN, B. J., A. SPINK et T. SARACEVIC 2000 «Real life, real users, and real needs: a study and analysis of user queries on the web», *Information Processing and Management* 36 : 207-227.

KROVETZ, R. 1993 «Viewing Morphology as an Inference Process», dans H.P., Frei et coll. *Proceedings of ACM-SIGIR 93*, pp. 191-202.

NAMER, F. 2000 «FLEM : un analyseur flexionnel du français à base de règles», *TAL* 41-2 : 523-547.

PEREZ-CARBALLO J. et T. STRZALKOWSKI 2000 «Natural language information retrieval: progress report», *Information Processing and Management* 36 : 155-178.

SPÄRCK Jones K. 1999 «The role of NLP in text retrieval», dans Strzalkowski, T. et coll. 1999 pp. 1-24.

STRZALKOWSKI, T. 1995 «Natural language information retrieval», *Information Processing and Management*, 31-3 : 397-417.

STRZALKOWSKI, T. et coll. 1999 *Natural Language Information Retrieval*, Dordrecht, Kluwer.

WOODS, W.A., L.A. BOOKMAN, A. HOUSTON, R.J. KUHNS, P. MARTIN, S. GREEN et coll. 2000 «Linguistic Knowledge can improve information retrieval», *Processing of the 6th Applied Natural Language Processing Conference*.